

Alex Teboul · Updated a year ago

A picture containing clipart

Description automatically generated

**Diabetes Health Indicators Dataset**

253,680 survey responses from cleaned BRFSS 2015 + balanced dataset



**About Dataset**

**Context**

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk. While there are different types of diabetes, type II diabetes is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health. Much of the burden of the disease falls on those of lower socioeconomic status as well. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approaching $400 billion dollars annually.

**Content**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

This dataset contains 3 files:

1. diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables
2. diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.
3. diabetes \_ binary \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

Explore some of the following research questions:

1. Can survey questions from the BRFSS provide accurate predictions of whether an individual has diabetes?
2. What risk factors are most predictive of diabetes risk?
3. Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?
4. Can we create a short form of questions from the BRFSS using feature selection to accurately predict if someone might have diabetes or is at high risk of diabetes?

**Acknowledgements**

It it important to reiterate that I did not create this dataset, it is just a cleaned and consolidated dataset created from the BRFSS 2015 dataset already on Kaggle. That dataset can be found [here](https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system) and the notebook I used for the data cleaning can be found [here](https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset-notebook).

**Inspiration**

Zidian Xie et al for Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques using the 2014 BRFSS was the inspiration for creating this dataset and exploring the BRFSS in general. [Link](https://www.cdc.gov/pcd/issues/2019/19_0109.htm)

**Capstone Two: Project Proposal**

**Problem Identification**

**Problem Statement**

A healthcare organization wants to predict with 90% accuracy if a person might have diabetes or is at risk of diabetes by finding out what risk factors are most predictive of diabetic risk.

**Context**

Diabetes is one of the most prominent chronic diseases in the United States. Early diagnosis can lead to lifestyle changes and more effective treatment. Having a predictive model for diabetic risk would be a valuable tool for the public and public health officials. This healthcare organization wants to make an impact in the healthcare field by being able to predict if someone might have diabetes or is at risk of diabetes by finding out what risk factors are the most predictive of diabetic risk.

**Criteria for Success**

A healthcare organization wants to predict with 90% accuracy if a person might have diabetes or is at risk of diabetes and find out what risk factors are most predictive of diabetic risk. There is no cure for this medical condition, but the early diagnosis of diabetes can help mitigate the harms of this disease in many patients and increase life expectancy.

**Scope of Solution Space**

The healthcare organization wants to predict with 90% accuracy if a person might have diabetes or is at risk of diabetes. This predictive model not only helps in early diagnosis but also increases the quality of life and decreases the financial burden on our economy.

**Constraints within Solution Space**

The dataset is from the BRFSS 2015

Datasets are survey questions from BRFSS

**Stakeholders to provide Key Insight**

Healthcare organization upper management

**Key data sources**

The source of data is Kaggle which has 3 data files cleaned and a usability score of 10.

This dataset has 21 feature variables and is balanced and is created from the BRFSS 2015 dataset.

**Problem Method and Solution**

* Identification of the problem:
  + Write a problem statement based on the client’s needs.
* Data wrangling:
  + Collect, organize, and clean the dataset: data acquisition, data shaping, and data quality.
  + Tidy data, reshaping data by sorting, reindexing, renaming, subsets rows, and columns, using a query, summarizing data, handling missing data, making new columns, combining data sets, grouping data, and plotting data.
* Perform exploratory data analysis:
  + Assess data quality, plot different combinations of variables, data visualization (point plots, scatter plots, bar plots, histogram, or line plots) depending on what you are communicating in the plot, profile the data, and explore the data across many dimensions.
* Preprocess the data by standardizing and training the dataset:
  + Imputing missing values, transforming the values, encoding categorical variables, train-test split method, measuring metric performance, cross-validation of data
* Modeling the dataset by training the model to make predictive insights:
  + Selection of the best algorithm for the problem, performing inference for modeling behavior
* Documentation of the work done:
  + Storytelling using a slide deck, and written report.
* Deliverables:
  + Writing codes, writing project report summary with results, and presentation to stakeholders using slide deck format